PB83-209825

Statistical Methods for Estimating Attributable Risk from Retrospective Data

Stanford Univ., CA

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1982

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before comple

1. REPORT NO.	2.	3.
EPA-600/J-82-317	JOURNAL ARTICLE	PB83-209825

4. TITLE AND SUBTITLE	5. REPORT DATE
Statistical Methods for Estimating Attributable Risk from Retrospective Data	1982
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Alice S. Whittemore	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Stanford University, School of Medicine Stanford, California 94305	11. CONTRACT/GRANT NO.
	CR806572
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Health Effects Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711	14. SPONSORING AGENCY CODE EPA/600/11

15. SUPPLEMENTARY NOTES

Statistics in Medicine, 1:229-243 (1982)

16. ABSTRACT

This paper extends Levin's measure of attributable risk[1] to adjust for confounding by aetiologic factors other than the exposure of interest. One can estimate this extended measure from case-control data provided either (i) from the control data one can estimate exposure prevalence within each stratum of the confounding factor; or (ii) one has additional information available concerning the confounder distribution and the stratum-specific disease rates. In both cases we give maximum likelihood estimates and their estimated asymptotic variances, and show them to be independent of the sampling design (matched vs. random). Computer simulations investigate the behaviour of these estimates and of three types of confidence intervals when sample size is small relative to the number of confounder strata. The simulations indicate that attributable risk estimates tend to be too low. The bias is not serious except when exposure prevalence is high among controls. In this case the estimates and their standard error estimates are also highly unstable. In general, the asymptotic standard error estimates performed quite well, even in small samples, and even when the true asymptotic standard error was too small. By contrast, the bootstrap estimate[2] tended to be too large. None of the three confidence intervals proved superior in accuracy to the other two. Thus there appears no advantage in using the log-based interval suggested by Walter[3,4] which is always longer than the simpler symmetric interval.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Attributable risk, Case-control studies, Maximum likelihood estimates, Small sample behavior		

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS *(This Report)*	21. NO. OF PAGES
Release to Public	Unclassified	16
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA-600/J-82-317
JOURNAL ARTICLE

STATISTICS IN MEDICINE, VOL. 1, 229–243 (1982)

STATISTICAL METHODS FOR ESTIMATING ATTRIBUTABLE RISK FROM RETROSPECTIVE DATA

ALICE S. WHITTEMORE

Department of Family, Community and Preventive Medicine, Stanford University School of Medicine, Stanford, California 94305, U.S.A.

SUMMARY

This paper extends Levin's measure of attributable risk[1] to adjust for confounding by aetiologic factors other than the exposure of interest. One can estimate this extended measure from case-control data provided either (i) from the control data one can estimate exposure prevalence within each stratum of the confounding factor; or (ii) one has additional information available concerning the confounder distribution and the stratum-specific disease rates. In both cases we give maximum likelihood estimates and their estimated asymptotic variances, and show them to be independent of the sampling design (matched vs. random). Computer simulations investigate the behaviour of these estimates and of three types of confidence intervals when sample size is small relative to the number of confounder strata. The simulations indicate that attributable risk estimates tend to be too low. The bias is not serious except when exposure prevalence is high among controls. In this case the estimates and their standard error estimates are also highly unstable. In general, the asymptotic standard error estimates performed quite well, even in small samples, and even when the true asymptotic standard error was too small. By contrast, the bootstrap estimate[2] tended to be too large. None of the three confidence intervals proved superior in accuracy to the other two. Thus there appears no advantage in using the log-based interval suggested by Walter[3,4] which is always longer than the simpler symmetric interval.

KEY WORDS Attributable risk Case-control studies Maximum likelihood estimates Small sample behaviour

INTRODUCTION

To evaluate disease prevention strategies, one must estimate the fraction of disease that might be avoided by reduction or elimination of population exposure to a given aetiologic agent. Relative risk and odds-ratio measures are inadequate for this purpose. Although the relative risk of pancreatic cancer due to familial relapsing pancreatitis is substantial,[3] this hereditary condition is so rare that its elimination would prevent only a small fraction of pancreatic cancers. By contrast, the relative risk of pancreatic cancer associated with cigarette smoking is only about 2;[4] nevertheless, elimination of this ubiquitous exposure would prevent a sizeable fraction of the disease.

Thus it is useful to have a measure of the proportion of disease attributable to a given exposure, one that accounts for correlations between the exposure and other aetiologic factors. Levin's simple measure of attributable risk,[1,5,6] first proposed in 1953, does not account for such confounding. The purpose of this paper is (i) to extend Levin's definition to account for confounding by other aetiologic factors; (ii) to present point estimates and confidence intervals for this extended measure based on matched or randomly sampled case-control data; and (iii) to investigate, using simulated data, the small sample properties of these estimates and confidence

Received December 1981
Revised June 1982

intervals. The extension of Levin's measure, defined below, has been described independently by Walter.[7]

DEFINITIONS AND THEORY

Consider a population that is classified into two groups according to the presence or absence of some exposure characteristic E. Let $P(D|E)$, $P(D|\bar{E})$ and $P(D)$ represent the incidence (or prevalence) rates of disease among the exposed, the unexposed, and the entire population, respectively. We assume that $P(D|E) \geq P(D|\bar{E})$. The ratio

$$P(D|\bar{E})/P(D)$$

represents the fraction of all disease that would persist if the entire population contracted disease at the rate of those unexposed to E; we label this ratio the *residual risk* corresponding to E. The remaining fraction,

$$1 - P(D|\bar{E})/P(D) \tag{1}$$

represents the proportion of all disease that one could prevent if the rates for the unexposed prevailed in the population. Levin[1] called this fraction the (*population*) *risk attributable to E*.

If exposure is correlated with an aetiologic factor C, then elimination of E from the population need not produce the unexposed disease rate $P(D|\bar{E})$. Instead, the disease rate in the absence of E would be a weighted average of stratum-specific unexposed rates. Algebraically, this rate is

$$\Sigma_k P(C_k)P(D|\bar{E}C_k) \tag{2}$$

where $P(C_k)$ represents the fraction of the population in the kth stratum of C, and $P(D|\bar{E}C_k)$ is the stratum-specific disease rate among the unexposed. Substitution of (2) for $P(D|\bar{E})$ in (1) yields a measure of attributable risk which has been adjusted for C:

$$AR = 1 - [\Sigma_k P(C_k)P(D|\bar{E}C_k)/P(D)] \tag{3}$$

Note that when C has only one level (3) reduces to (1). By using in (1) the identity

$$P(D|\bar{E}) = \Sigma_k P(D|\bar{E}C_k)P(C_k|\bar{E})$$

where the quantities $P(C_k|\bar{E})$ represent the confounder distribution among the unexposed, one can express the difference between the unadjusted measure (1) and the adjusted measure (3) as

$$\{\Sigma_k [P(C_k) - P(C_k|\bar{E})]P(D|\bar{E}C_k)\}/P(D) \tag{4}$$

Expression (4) indicates that either of the following is a sufficient condition for agreement between adjusted and unadjusted measures of attributable risk: (i) C is unrelated to disease among the unexposed, so that $P(D|\bar{E}C_k) = P(D|\bar{E})$ for all strata C_k; or (ii) E and C are uncorrelated, so that $P(C_k|\bar{E}) = P(C_k)$ for all k. These conditions, noted by Walter,[7] are also necessary when C has only two levels. They imply that one must control confounding factors in measuring risk attributable to E.

ESTIMATION

We can rewrite the attributable risk AR in (3) as

$$AR = \Sigma_k P(C_k|D)AR_k \tag{5}$$

where

$$AR_k = 1 - [P(D|\bar{E}C_k)/P(D|C_k)] = 1 - [P(\bar{E}|C_k D)/P(\bar{E}|C_k)] \tag{6}$$

is the stratum-specific attributable risk. To estimate AR solely from case-control data one needs the 'rare disease assumption' that within each stratum C_k the proportion $P(\overline{E}|C_k)$ of unexposed can be approximated by the corresponding proportion $P(\overline{E}|C_k\overline{D})$ among the non-diseased. If so, then from (5) and (6) we have approximately

$$AR = 1 - \Sigma_k P(C_k|D)P(\overline{E}|C_kD)/P(\overline{E}|C_k\overline{D}) \tag{7}$$

a parameter estimable from case-control data. Equation (7) holds only in the limit as the probability of disease tends to zero, a distinction that we will ignore. When C has only one level, (7) reduces to the unadjusted measure

$$1 - [P(\overline{E}|D)/P(\overline{E}|\overline{D})] \tag{8}$$

Random Sampling

The investigator using a random sampling design chooses random samples consisting of n diseased subjects (cases) and m disease-free subjects (controls). He then can arrange his case and control exposure and confounder data into the $2 \times 2 \times K$ Table I. The $2K$ dimensional vectors $\mathbf{x} = (x_{11}, \ldots, x_{1K}, x_{21}, \ldots, x_{2K})$ and $\mathbf{y} = (y_{11}, \ldots, y_{1K}, y_{21}, \ldots, y_{2K})$ of case and control frequencies each have a multinomial distribution, with respective parameters n, $P(EC_k|D)$, $P(\overline{E}C_k|D)$ and m, $P(EC_k|\overline{D})$, $P(\overline{E}C_k|\overline{D})$, $k = 1, \ldots, K$; moreover $\mathbf{x}$ and $\mathbf{y}$ are independent. The maximum likelihood estimate (MLE) of AR in (7) is

$$\widehat{AR} = 1 - (\Sigma_k y_{\cdot k} x_{2k}/n y_{2k}) \tag{9}$$

Table I. Numbers of cases and controls by exposure status and confounder stratum

Disease status	Exposure status	Confounder stratum			
		C_1	C_2	C_k	C_K
cases	Exposed	x_{11}	x_{12}	x_{1k}	x_{1K}
	Unexposed	x_{21}	x_{22}	x_{2k}	x_{2K}
controls	Exposed	y_{11}	y_{12}	y_{1k}	y_{1K}
	Unexposed	y_{21}	y_{22}	y_{2k}	y_{2K}

where the dot subscript indicates summation. Standard large sample methods show that as m and n increase, $\widehat{AR}$ becomes normally distributed with asymptotic mean AR and with variance given by expression (22) in Appendix I. A consistent estimator for this variance is

$$u^2 = \frac{1}{n^2}\left[\Sigma_k (y_{\cdot k} x_{2k}/y_{2k})^2 \left(\frac{1}{x_{2k}} + \frac{y_{1k}}{y_{\cdot k} y_{2k}}\right) - n(1 - \widehat{AR})^2\right] \tag{10}$$

Relatively accurate population estimates may exist both for the stratum-specific disease rates $\phi_k = P(D|C_k)$ and for the confounder distribution $P(C_k)$. If these probabilities are known, one can estimate AR from case-control data without the rare disease assumption. To do so, rewrite the stratum-specific attributable risks AR_k of (6) as

$$AR_k = 1 - [\phi_k + (1 - \phi_k)P(\overline{E}|C_k\overline{D})/P(\overline{E}|C_kD)]^{-1} \tag{11}$$

The MLE of AR as given by (5) and (11) is

$$\widetilde{AR} = \Sigma \sigma_k \widetilde{AR}_k \tag{12}$$

In (12), σ_k is the known parameter $P(C_k|D) = \phi_k P(C_k)/\Sigma_k \phi_k P(C_k)$, and $\widetilde{AR}_k$ results from substitution in (11) of sample moments $x_{2k}/x_{\cdot k}$ and $y_{2k}/y_{\cdot k}$ for the unknown probabilities $P(\overline{E}|C_k D)$ and $P(\overline{E}|C_k \overline{D})$. The large sample variance of $\widetilde{AR}$ appears in the appendix as equations (28) and (29). A consistent estimate of this variance is

$$v^2 = \Sigma_k [(1 - \widetilde{AR}_k)^2 \sigma_k (1 - \phi_k)(y_{2k} x_{\cdot k}/x_{2k} y_{\cdot k})]^2 \left[\frac{x_{1k}}{x_{\cdot k} x_{2k}} + \frac{y_{1k}}{y_{\cdot k} y_{2k}}\right] \tag{13}$$

The estimates $\widehat{AR}$, $\widetilde{AR}$, u^2 and v^2 all reduce to the corresponding unadjusted expressions described by Walter[5] when $K = 1$.

Matched sampling

The investigator using a matched design obtains information on the variables E and C for each of n randomly sampled cases. Again, he can arrange these data as in Table I. The vector $\mathbf{x}$ of case frequencies has the multinomial distribution described for the random sampling design. Corresponding to the $x_{\cdot k}$ cases in stratum C_k the investigator randomly chooses $R_k x_{\cdot k}$ controls from the same stratum, where R_k need not be an integer. For this design the control frequency vector $\mathbf{y}$ of Table I is not independent of $\mathbf{x}$. Instead, given the $x_{\cdot k}$, the y_{1k} are independent binomial variates with parameters $R_k x_{\cdot k}$ and $P(E|\overline{D}C_k)$, $k = 1, \ldots, K$.

Equation (7) shows that under the rare disease assumption, one can estimate AR from data obtained in the matched design; equation (9) again gives the MLE. One also sees that the pooled estimate $1 - (\Sigma_k R_k x_{\cdot k}) x_{\cdot 2}/(n y_{\cdot 2})$ for AR is inconsistent. Indeed, the pooled estimate converges in probability to the expression

$$1 - [P(\overline{E}|D)\Sigma_k R_k P(C_k|D)/\Sigma_k R_k P(C_k|D) P(\overline{E}|C_k \overline{D})] \tag{14}$$

which, in general, equals neither (7) nor (8). When the control: case ratio R_k is the same for all strata, (14) reduces to

$$1 - [P(\overline{E}|D)/\Sigma_k P(\overline{E}|C_k \overline{D})] \tag{15}$$

Sufficient conditions for equality between (15) and (7) are either (i) the factor C is independent of disease among both exposed and unexposed; or (ii) C is independent of exposure among the controls. In practice, neither of these conditions will likely hold, for, if so, matching on C is unnecessary.

As n becomes large the distribution for $\widehat{AR}$ becomes normal with mean AR and with variance given by appendix equation (26), an expression almost identical to the asymptotic variance (22) corresponding to the unmatched design. Use of sample moments for the unknown parameters in (26) again yields the consistent estimator u^2 of (10).

When one assumes known stratum-specific disease rates and known confounder distribution, the MLE for AR obtained from the matched design is again $\widetilde{AR}$ of (13), with asymptotic variance given by appendix equations (30) and (28), and consistently estimated by v^2 of (13).

Thus both matched and random sampling designs produce the same point estimates $\widehat{AR}$ and $\widetilde{AR}$, and the same variance estimates u^2 and v^2.

Confidence intervals

An approximate $100(1-\alpha)$ per cent confidence interval for attributable risk based on the maximum likelihood methods of the preceding sections is

$$C_{ML}: (\widehat{AR} - w,\ \widehat{AR} + w) \tag{16}$$

where

$$w = z_{1-\alpha/2}u \tag{17}$$

Investigators have proposed various monotonic transformations, $g(AR)$, to produce confidence intervals for attributable risk with smaller length or more accurate coverage probability. The log transformation, $g(AR) = \ln(1 - AR)$, suggested by Walter[5,6] gives rise to the $100(1-\alpha)$ per cent confidence interval

$$C_{LOG}: (1 - \hat{\rho}e^{w/\hat{\rho}},\ 1 - \hat{\rho}e^{-w/\hat{\rho}}) \tag{18}$$

where $\hat{\rho} = 1 - \widehat{AR}$ is the residual risk estimate. The logit transformation, $g(AR) = \ln[AR/(1-AR)]$ $(0 < AR < 1)$, proposed by Leung and Kupper,[8] yields

$$C_{LOGIT}: \left(\left[1 + (\hat{\rho}/\widehat{AR})e^{w/(\hat{\rho}\widehat{AR})}\right]^{-1},\ \left[1 + (\hat{\rho}/\widehat{AR})e^{-w/(\hat{\rho}\widehat{AR})}\right]^{-1}\right) \tag{19}$$

One may derive confidence intervals for use when one knows both the stratum specific disease probabilities and confounder distribution by replacing $\widehat{AR}$ by $\widetilde{AR}$ and u by v in (16)–(19).

Appendix II shows that the ML interval is always shorter than the log interval (except for the trivial case when both intervals have length zero). Leung and Kupper[8] have also shown that the logit interval is shorter than the ML interval for attributable risk estimates between 21 per cent and 79 per cent, whereas the reverse tends to hold for estimates less than 21 per cent.

EXAMPLE

Table II gives lung cancer mortality data among white U.S. uranium miners by cigarette smoking and radiation exposure status. Controls consisted of 776 miners matched to the 194 lung cancer decendents on year of birth in the ratio 4 : 1. Cumulative radon-daughter exposure measured in working level months (WLM) was the exposure of interest; E is exposure $\geqslant 120$ WLM. One can find details on exposure calculations in Reference 9, and on the case-control analysis of this uranium mining cohort in Reference 10.

In estimating attributable risk, we assigned the value of 0·5 to the zero cells in the table. Assuming that the controls yield an estimate of radiation exposure prevalence among uranium miners, one finds from the study data alone that the age-adjusted attributable risk is $\widehat{AR} = 52{\cdot}5\%$ with estimated standard error $u = 7{\cdot}9\%$. Age-adjusted 95 per cent confidence intervals for AR are $C_{LOGIT} = (37{\cdot}2, 67{\cdot}3)$; $C_{ML} = (37{\cdot}1, 68{\cdot}0)$; and $C_{LOG} = (34{\cdot}3, 65{\cdot}7)$. The attributable risk estimate adjusted both for age (matching factor) and cigarette smoking (not a matching factor) is 51·5 per cent, with $u = 8{\cdot}2$ per cent. The unadjusted (inconsistent) attributable risk estimate is 53·9 per cent with $u = 2{\cdot}0$ per cent. For these data adjusted and unadjusted estimates are in close agreement, and the unadjusted estimate has smaller standard error.

One can also estimate age-adjusted attributable risk by supplementing the case-control data with 1970 census data[11] and with white male lung cancer mortality rates,[12] as shown in Table II. This yields the estimate $\widetilde{AR} = 52{\cdot}9\%$ with $v = 10{\cdot}8$ per cent, in good agreement with the corresponding age-adjusted estimates obtained solely from the case-control data.

SIMULATION RESULTS

We use simulated data in this section to examine the small sample behaviour of $\widehat{AR}$ and u and of confidence intervals based on log, ML and logit procedures. We undertook two simulation studies. In the first, we conducted 1000 trials (i.e. 'case-control studies') for each of 72 combinations of parameter values, numbers of strata and sample sizes, using both matched and unmatched designs,

Table II. Exposure data for U.S. uranium miners and mortality rates for lung cancer among U.S. white males

Age at case's death (years)		Number of white miners cumulative radiation exposure (WLM*) < 120 Cumulative smoking (pack-years†) < 20	< 120 ≧ 20	≧ 120 < 20	≧ 120 ≧ 20	1977 U.S. white male lung cancer mortality rate (per 10^5)[11]	1970 U.S. white male population in thousands[12]	σ_k‡
30–44	Cases	3	0	12	1	9·2	14903·9	0·026
	Controls	21	1	39	3			
45–49	Cases	4	3	12	12	49·5	5257·6	0·049
	Controls	39	18	42	25			
50–54	Cases	4	2	10	12	97·9	4832·6	0·090
	Controls	29	27	26	30			
55–59	Cases	2	5	5	34	160·4	4310·9	0·131
	Controls	35	44	41	64			
60–64	Cases	1	5	3	24	266·1	3647·2	0·184
	Controls	18	22	35	57			
65–69	Cases	1	3	3	9	359·5	2808·0	0·193
	Controls	7	11	16	30			
70–79	Cases	0	1	2	21	485·0	3545·2	0·327
	Controls	13	17	35	31			
Total	Cases	15	19	47	113	138·7	39305·4	1·000
	Controls	162	140	234	240			

* 1 WLM = 1 WL × 170 hours, where 1 WL = $1{\cdot}3 \times 10^5$ Mev of potential alpha energy per litre of air.
† One pack-year equals 365 packs of cigarettes.
‡ σ_k is the product of column 5 and column 6, normalized so that $\Sigma_k \sigma_k = 1$.

a total of $72 \times 2 \times 1000 = 144{,}000$ trials. For each trial of unmatched design, we generated two independent samples of n and m independent multinomial random variables. For each trial of matched design, we generated n independent multinomial random variables; corresponding to each such 'case' variable, we generated R independent Bernoulli 'control' variables with exposure probability determined by the stratum of the case variable.

For simplicity, we took for the stratum-specific exposure probabilities among cases and controls:

$$\begin{aligned} P(EC_k|D) &\equiv \pi/K, \qquad P(\overline{EC}_k|D) \equiv (1-\pi)/K \\ P(EC_k|\overline{D}) &\equiv p/K, \qquad P(\overline{EC}_k|\overline{D}) \equiv (1-p)/K \end{aligned} \tag{20}$$

We choose this model with identical true adjusted and unadjusted values of AR to minimize the number of adjustable parameters. Of course, since C is not a confounder, matching or stratification is unnecessary. We also investigated a model in which C is a confounding factor. Since the two models yielded similar results, we have omitted discussion of the latter investigation. We considered all possible combinations of the following sample sizes, numbers of strata and parameter values: $n = 100, 500$; $K = 1, 5, 10$; control : case ratio $R (= m/n) = 1, 3$; $\pi = 0{\cdot}2, 0{\cdot}5, 0{\cdot}8$; and $p = 0{\cdot}2, 0{\cdot}5, 0{\cdot}8$ (subject to $p \leqslant \pi$). When a cell frequency was zero, we assigned it a value of 0·5.

Table III shows average values of attributable risk $\widehat{AR}$ for the matched designs. Corresponding results (not shown) for the unmatched designs were similar. Table III indicates that attributable risk estimates are biased downward; this bias increases with increasing number of strata, with

Table III. Attributable risk estimates (‰) averaged over 1000 matched case-control studies (simulated data)

$P(E\|D)$ (‰)	$P(E\|\overline{D})$ (‰)	AR (‰)	Number of strata	Number of cases: 100, Case : control ratio 1 : 1	100, 1 : 3	500, 1 : 1	500, 1 : 3
20	20	0	1	−0·5	0·0	−0·1	−0·1
			5	−1·6	−0·4	−0·3	−0·1
			10	−3·0	−0·9	−0·6	−0·2
50	20	37·5	1	37·9	37·0	37·5	37·5
			5	36·7	37·0	37·3	37·3
			10	35·6	36·9	37·3	37·6
	50	0	1	−1·0	−0·1	−0·5	−0·3
			5	−5·4	−2·6	−1·1	−0·2
			10	−14·8	−4·1	−2·1	−0·4
80	20	75	1	74·8	74·8	75·0	75·0
			5	74·4	74·9	74·9	75·0
			10	73·3	73·9	74·9	75·0
	50	60	1	59·2	59·5	59·6	59·8
			5	58·2	59·1	59·6	59·8
			10	52·3	57·0	59·0	59·8
	80	0	1	−4·7	−7·7	−1·1	−0·2
			5	−35·2	−9·8	−4·8	−1·1
			10	−48·6	−23·7	−10·0	−2·2

increasing values of π and p, and with decreasing sample size. The bias is unusually severe for the case $\pi = p = 80$ per cent. This suggests that attributable risk estimates are unreliable when exposure prevalence is high among controls.

The bias is not surprising in view of the inconsistency of the maximum likelihood estimator for the common odds-ratio in several 2×2 tables, as the number of tables increases while the data within each table remain sparse.[13] Indeed, one can show that in the extreme example of matched studies with one case and one control per stratum, the estimator $\hat{AR}$ for a true AR of zero converges in probability to a negative number.

We also calculated observed confidence coefficients and average lengths for each of the three types of 95 per cent confidence interval. We did not calculate the logit interval in trials for which the true AR was zero, because the interval is undefined when $\hat{AR} \leqq 0$, an event which occurred in more than half of such trials. For trials involving 500 cases, all observed coefficients were between 94 per cent and 96 per cent, and none of the three methods proved uniformly more accurate than the others. The average lengths of the intervals differed from one another by no more than 0·3 % and increased with the number of strata. The increase was slight except for the case $\pi = p = 80$ per cent, $R = 1$, for which the average length increased from 50·4 per cent ($K = 1$) to 66·5 per cent ($K = 10$).

Tables IV and V display observed coefficients and average lengths for trials with 100 cases. The tables show that, for all three types of intervals, average length and extent of disagreement between observed and nominal coefficients increase with both exposure prevalence and number of strata. In the case where $\pi = p = 80$ per cent the ML and log intervals are conservatively long. According to Table IV, no one method is clearly superior in achieving the nominal coefficient. As expected, Table V shows a consistent gradient in length among the three types of interval, with the logit interval shortest and the log interval longest.

We designed a second simulation study to investigate the accuracy of u and Efron's bootstrap[2] as estimates of the standard error of $\hat{AR}$ when sample size is small relative to number of confounder strata. We conducted ten trials or 'case-control studies' for each combination of the parameter values, strata numbers, sample sizes and study designs described previously, a total of 1440 trials. Each trial yielded two estimates of standard error: (i) u of (10); (ii) a bootstrap estimate, BSE. (Appendix III describes the computations to obtain BSE.)

We compared values of u and BSE, averaged over the ten trials, with the standard deviation SD of 1000 attributable risk estimates, defined by

$$SD = \left[\frac{1}{999} \sum_{i=1}^{1000} (\hat{AR}_i - \overline{AR})^2 \right]^{1/2} \tag{21}$$

where $\overline{AR}$ is the average of the 1000 estimates. We also compared SD with the true asymptotic standard error, (23) or (28), which, for the simple model (20), depends neither on K nor on the study design. Except for the problem case $\pi = p = 80$ per cent, u agreed well with SD, even for trials involving only 100 cases and 100 controls. By contrast, the bootstrap estimate tended to be larger than SD, at times substantially so, whereas the asymptotic standard error was often too small. These discrepancies were quite pronounced in the troublesome case $\pi = p = 80$ per cent, as shown in Table VI. The large and fluctuating standard errors shown in the table provide further evidence of the instability of attributable risk estimates when exposure is common among both cases and controls. The instability is understood upon examination of $\hat{AR}$ in (9). When exposure is common among controls, the denominators y_{2k} have probability close to zero, and in small samples these variables will often be zero (in practice, 0·5), leading to very large values for the expression in brackets.

Table IV. Observed coefficients (per cent) of 95 per cent confidence intervals for AR using 1000 case-control simulations; matched design; $n = 100$ cases

$P(E\|D)$ (%)	$P(E\|\bar{D})$ (%)	AR (%)	Case: control ratio	Number of Strata								
				1			5			10		
				LOGIT	ML	LOG	LOGIT	ML	LOG	LOGIT	ML	LOG
20	20	0	1:1	–*	94·4	94·2	–	95·8	95·6	–	96·9	96·5
			1:3	–	94·9	94·8	–	94·7	94·8	–	94·5	94·4
50	20	37·5	1:1	96·5	95·2	95·8	97·0	94·6	94·4	97·8	95·8	94·6
			1:3	96·1	94·7	93·5	95·6	94·8	96·0	96·9	95·1	94·4
	50	0	1:1	–	95·6	95·9	–	96·7	96·5	–	98·4	96·3
			1:3	–	94·7	94·7	–	96·6	95·7	–	97·1	95·2
80	20	75	1:1	95·8	94·4	95·0	96·2	94·6	95·1	96·0	96·6	94·7
			1:3	95·1	94·7	94·6	95·3	93·8	94·8	97·4	97·4	96·4
	50	60	1:1	96·4	93·8	94·3	97·7	96·2	96·4	99·3	98·6	96·0
			1:3	94·7	93·1	96·1	96·8	95·5	95·0	98·3	96·6	95·3
	80	0	1:1	–	94·1	95·6	–	98·7	98·7	–	99·7	99·1
			1:3	–	95·3	94·3	–	95·7	94·7	–	99·3	96·1

* Not calculated (see text)

Table V. Average length (per cent) of 95 per cent confidence intervals for AR using 1000 case-control simulations: matched design; $n = 100$ cases

$P(E\|D)$ (%)	$P(E\|\bar{D})$ (%)	AR (%)	Case: control ratio	Number of strata: 1			5			10		
				LOGIT	ML	LOG	LOGIT	ML	LOG	LOGIT	ML	LOG
20	20	0	1:1	–*	28·0	28·1	–	30·0	30·1	–	33·6	33·8
			1:3	–	22·6	22·6	–	23·0	23·0	–	23·6	23·7
50	20	37·5	1:1	25·0	25·4	25.6	28·0	28·6	28.9	30·4	31·2	31·5
			1:3	12·2	12·2	12·3	25·2	25·7	25·9	25·6	26·1	26·3
	50	0	1:1	–	56·8	57·5	–	66·8	68·0	–	100·2	107·1
			1:3	–	45·3	45·7	–	47·6	48·0	–	50·8	51·3
80	20	75	1:1	20·0	20·1	20·7	20·7	20·9	21·5	22·0	22·2	22·9
			1:3	19·6	19·7	20·2	19·7	19·8	20·3	20·2	20·3	20·9
	50	60	1:1	34·2	35·6	36·8	37·5	39·6	41·1	47·6	55·9	61·5
			1:3	31·6	32·7	33·6	32·5	33·7	34·7	34·4	35·9	37·1
	80	0	1:1	–	118·4	125·0	–	253·6	321·6	–	304·7	375·5
			1:3	–	92·1	95·4	–	110·2	115·2	–	156·3	173·2

* Not calculated (see text)

Table VI. Standard error (per cent) of attributable risk estimates: $P(E|D) = P(E|\bar{D}) = 80$ per cent, $AR = 0$ per cent

Number of cases	Case: control ratio	Asymptotic standard error*	Number of strata: 1			Number of strata: 5, Study design: Unmatched			Number of strata: 5, Study design: Matched			Number of strata: 10, Study design: Unmatched			Number of strata: 10, Study design: Matched		
			SD†	*u*‡	*BSE*§	*SD*	*u*	*BSE*	*SD*	*u*	*BSE*	*SD*	*u*	*BSE*	*SD*	*u*	*BSE*
100	1:1	28	32	33	39	63	75	81	51	67	68	51	73	57	51	67	68
	1:3	23	23	23	33	29	31	77	34	35	57	38	72	61	34	35	57
500	1:1	13	13	14	14	14	13	15	17	18	32	16	17	29	17	18	32
	1:3	10	10	10	13	11	11	15	11	11	11	11	11	12	11	11	11

* From (20) and either (22) or (26)
† Standard deviation of 1000 attributable risk estimates; see formula (21)
‡ Formula (10), averaged over 10 trials
§ Bootstrap estimate of standard error using 200 replications, averaged over 10 trials

CONCLUSIONS

The results of Walter[7] and of the preceding sections indicate the necessity of adjustment for confounding when estimating attributable risk. Adjusted estimates derive from case-control studies provided that one can estimate stratum-specific population exposure prevalence from control data. Estimation is still possible when this assumption fails, provided stratum-specific disease rates are available. In both cases, the investigator obtains the same estimates and confidence intervals for attributable risk when controlling confounding by matching as when controlling it by stratifying during analysis.

Attributable risk estimates obtained using the 2×2 table of pooled data from a matched study are inconsistent. Breslow[14] has noted that in practice the analogous odds-ratio estimation procedure involves small bias, and the bias may also be small in this case.

The simulation studies described in the preceding section show that in small samples the maximum likelihood estimate for attributable risk tends to be too low. The bias is usually small, provided the number of strata is not large relative to sample size. An exception occurs when exposure is common among controls. In this case the estimates have severe downward bias and large standard errors, particularly when sample size is small relative to the number of confounder strata. Confidence intervals are conservatively long, and standard error estimates are themselves unstable.

In general, the asymptotic standard error estimate showed good agreement with the standard deviation of $\widehat{AR}$ from the simulation studies, even when sample sizes were small. The bootstrap estimate was frequently too large, and the true asymptotic standard error too small.

For large samples the three types of confidence intervals differed little in their coverage probability and average length. Small samples often resulted in sizable differences in length; the log interval length exceeded that of the ML interval, and, when $0{\cdot}21 < \widehat{AR} < 0{\cdot}79$, the ML interval length exceeded that of the logit interval. A plot of the distributions of $\widehat{AR}$ obtained from 1000 trials involving 100 cases and 100 controls for selected parameter combinations displayed slight skewness to the left, whereas the distributions of $\ln(1 - \widehat{AR})$ indicated more symmetry. Although this suggests improved accuracy for the log interval, the simulation studies did not support this. Thus the longer log interval appears to have no advantage. Instead, one should use the logit interval for attributable risk estimates in the range 0·21 to 0·79, and the ML interval for attributable risk estimates outside this range.

ACKNOWLEDGEMENTS

This research was supported by NIH Grants Nos. ES 00070 and CA 23214, by NIOSH Grant No. OH 00991, and by grants to the SIAM Institute for Mathematics and Society from the Environmental Protection Agency, the Sloan Foundation, and the National Science Foundation.

APPENDIX I. ASYMPTOTIC VARIANCE OF ATTRIBUTABLE RISK ESTIMATES

Denote the probability, $P(EC_k|D)$, that a case falls in cell EC_k by π_{1k}, and set $\pi_{2k} = P(\overline{EC}_k|D)$, $k = 1, \ldots, K$. Let $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ represent the column vectors whose kth components are π_{1k} and π_{2k}, respectively, with $\boldsymbol{\pi}' = (\boldsymbol{\pi}_1', \boldsymbol{\pi}_2')$. Let $\mathbf{p}_1, \mathbf{p}_2$ and $\mathbf{p}$ represent the corresponding probability vectors for the controls. Similarly, let $\mathbf{x}_1' = (x_{11}, \ldots, x_{1K})$ and $\mathbf{x}_2' = (x_{21}, \ldots, x_{2K})$ be the sample vectors of cell frequencies for the cases, with $\mathbf{x}' = (\mathbf{x}_1', \mathbf{x}_2')$. The control frequency vectors $\mathbf{y}_1, \mathbf{y}_2$ and $\mathbf{y}$ have analogous definitions. We conveniently obtain the asymptotic variance of the residual risk estimate $\hat{\rho} = 1 - \widehat{AR}$, using the delta method.

For the *unmatched design*, we obtain the asymptotic variance of $\hat{\rho} = \frac{1}{n}\Sigma y_{\cdot k} x_{2k}/y_{2k}$ as the variance of its first order Taylor series expansion

$$\hat{\rho} \doteq \rho + \left(\frac{1}{n}\mathbf{x}_2 - \boldsymbol{\pi}_2\right)' \boldsymbol{\rho}_{x_2} + \left(\frac{1}{m}\mathbf{y} - \mathbf{p}\right)' \boldsymbol{\rho}_y$$

Here $\rho = \Sigma_k p_{\cdot k}\pi_{2k}/p_{2k}$, and the vectors $\boldsymbol{\rho}_{x_2}$ and $\boldsymbol{\rho}_y$ of first partials are $\boldsymbol{\rho}_{x_2} = \langle p_{\cdot k}/p_{2k} \rangle$ and $\boldsymbol{\rho}_y' = (\pi_{21}/p_{21}, \ldots, \pi_{2K}/p_{2K}, -\pi_{21}p_{11}/p_{21}^2, \ldots, -\pi_{2K}p_{1K}/p_{2K}^2)$. The resulting asymptotic variance is

$$A \text{ var } \hat{\rho} = \boldsymbol{\rho}_{x_2}'[D(\boldsymbol{\pi}_2) - \boldsymbol{\pi}_2\boldsymbol{\pi}_2']\boldsymbol{\rho}_{x_2}/n + \boldsymbol{\rho}_y'[D(\mathbf{p}) - \mathbf{p}\mathbf{p}']\boldsymbol{\rho}_y/m$$

where $D(\mathbf{z})$ is the diagonal matrix whose entries are the components of the vector $\mathbf{z}$. This gives

$$\begin{aligned} A \text{ var } \hat{\rho} &= \frac{1}{n}\left[\Sigma_k\left(\frac{\pi_{2k}p_{\cdot k}}{p_{2k}}\right)^2\left(\frac{1}{\pi_{2k}} + \frac{np_{1k}}{mp_{\cdot k}p_{2k}}\right) - \rho^2\right] \\ &= \frac{1}{n}\left\{\Sigma_k\left(\frac{P(\overline{E}C_k|D)}{P(\overline{E}|C_k\overline{D})}\right)^2\left[\frac{1}{P(\overline{E}C_k|D)} + \frac{nP(E|C_k\overline{D})}{mP(\overline{E}C_k|\overline{D})}\right] - (1 - AR)^2\right\} \end{aligned} \tag{22}$$

Substitution of sample moments for the unknown probabilities in (22) yields the expression u^2 of (10).

For the *matched design* the conditional distribution of y_{2k} given $R_k x_{\cdot k}$ is binomial with parameters $R_k x_{\cdot k}$ and $p_{2k}/p_{\cdot k}$. We approximate the variance of $\hat{\rho}$ by the variance of the first order expansion

$$\hat{\rho} \doteq \rho + \left(\frac{1}{n}\mathbf{x} - \boldsymbol{\pi}\right)'\boldsymbol{\rho}_x + (\langle y_{2k}/n - R_k\pi_{\cdot k}p_{2k}/p_{\cdot k}\rangle)'\boldsymbol{\rho}_{y_2} \tag{23}$$

where the vectors of first partials are $\boldsymbol{\rho}_x' = (\pi_{21}p_{\cdot 1}/p_{21}\pi_{\cdot 1}, \ldots, \pi_{2K}p_{\cdot K}/p_{2K}\pi_{\cdot K}, (1 + \pi_{21}/\pi_{\cdot 1})p_{\cdot 1}/p_{21}, \ldots, (1 + \pi_{2K}/\pi_{\cdot K})p_{\cdot K}/p_{2K})$, and $\boldsymbol{\rho}_{y_2}' = \langle -(\pi_{2k}/\pi_{\cdot k})(p_{\cdot k}/p_{2k})^2/R_k \rangle$. From (23) we have

$$A \text{ var } \hat{\rho} = n^{-2}(\boldsymbol{\rho}_x', \boldsymbol{\rho}_{y_2}')S(\boldsymbol{\rho}_x', \boldsymbol{\rho}_{y_2}') \tag{24}$$

where S is the $3K \times 3K$ covariance matrix of $(\mathbf{x}', \mathbf{y}_2')$. The entries of S derive from the joint moment generating function for $(\mathbf{x}', \mathbf{y}_2')$, given by

$$\begin{aligned} &E[\exp(\Sigma_{j,k}t_{jk}x_{jk} + \Sigma_k s_{2k}y_{2k})] \\ &\qquad = \left[\Sigma_k(\pi_{1k}e^{t_{1k}} + \pi_{2k}e^{t_{2k}})\left(\frac{p_{1k}}{p_{\cdot k}} + \frac{p_{2k}}{p_{\cdot k}}e^{s_{2k}}\right)^{R_k}\right]^n \end{aligned}$$

where $t_{2K} = 0$. This gives

$$n^{-1}S = \begin{pmatrix} D(\boldsymbol{\pi}) & D(\boldsymbol{\gamma}_1) \\ & D(\boldsymbol{\gamma}_2) \\ D(\boldsymbol{\gamma}_1)D(\boldsymbol{\gamma}_2) & D(\boldsymbol{\gamma} + \boldsymbol{\eta}) \end{pmatrix} - \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\gamma} \end{pmatrix}(\boldsymbol{\pi}'\boldsymbol{\gamma}') \tag{25}$$

where $\boldsymbol{\gamma}_j = \langle \pi_{jk}R_k p_{2k}/p_{\cdot k} \rangle$, $j = 1, 2$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2$, and $\boldsymbol{\eta} = \langle R_k(R_k - 1)\pi_{\cdot k}(p_{2k}/p_{\cdot k})^2 \rangle$. Straightforward (and tedious) computation from (24) and (25) yields

$$A \operatorname{var} \hat{\rho} = \frac{1}{n}\left[\Sigma_k\left(\frac{\pi_{2k}p_{\cdot k}}{p_{2k}}\right)^2\left(\frac{1}{\pi_{2k}}+\frac{p_{1k}}{R_k\pi_{\cdot k}p_{2k}}\right)-\rho^2\right]$$

$$= \frac{1}{n}\left\{\Sigma_k\left(\frac{P(\bar{E}C_k|D)}{P(\bar{E}|C_k\bar{D})}\right)^2\left[\frac{1}{P(EC_k|D)}+\frac{P(E|C_k\bar{D})}{R_kP(C_k|D)P(\bar{E}|C_k\bar{D})}\right]-(1-AR)^2\right\} \tag{26}$$

Substitution of sample moments for the unknown parameters in (26) yields u^2 of (10).

Next, we calculate the asymptotic variance of $\tilde{\rho} = 1 - A\tilde{R} = \Sigma\sigma_k[\phi_k + (1-\phi_k)y_{2k}x_{\cdot k}/x_{2k}y_{\cdot k}]^{-1}$, the residual risk estimate obtained using known stratum-specific disease probabilities $\phi_k = P(D|C_k)$ and confounder distribution $\sigma_k = P(C_k|D)$. For the *unmatched design* we use the first order expansion

$$\tilde{\rho} \doteq \rho + \left(\frac{1}{n}\mathbf{x}-\boldsymbol{\pi}\right)'\boldsymbol{\rho}_x + \left(\frac{1}{m}\mathbf{y}-\mathbf{p}\right)'\boldsymbol{\rho}_y \tag{27}$$

where $\boldsymbol{\rho}_x' = (\boldsymbol{\rho}_{x_1}', \boldsymbol{\rho}_{x_2}')$ and $\boldsymbol{\rho}_y' = (\boldsymbol{\rho}_{y_1}', \boldsymbol{\rho}_{y_2}')$, and where the kth components, respectively, of $\boldsymbol{\rho}_{x_1}$, $\boldsymbol{\rho}_{x_2}$, $\boldsymbol{\rho}_{y_1}$, $\boldsymbol{\rho}_{y_2}$ are $-\tau_k/\pi_{\cdot k}$, $\tau_k\pi_{1k}/(\pi_{2k}\pi_{\cdot k})$, $\tau_k/p_{\cdot k}$, and $-\tau_k p_{1k}/(p_{2k}p_{\cdot k})$. Here

$$\tau_k = (1-AR_k)^2\sigma_k(\phi_k)P(\bar{E}|C_k\bar{D})/P(\bar{E}|C_kD), \tag{28}$$

with AR_k given by (11). The approximation (27) yields

$$A \operatorname{var} \tilde{\rho} = \boldsymbol{\rho}_x'(D(\boldsymbol{\pi})-\boldsymbol{\pi}\boldsymbol{\pi}')\boldsymbol{\rho}_x/n + \boldsymbol{\rho}_y'(D(\mathbf{p})-\mathbf{p}\mathbf{p}')\boldsymbol{\rho}_y/m$$

$$= \Sigma_k\tau_k^2\left(\frac{\pi_{1k}}{n\pi_{\cdot k}\pi_{2k}}+\frac{p_{1k}}{mp_{\cdot k}p_{2k}}\right)$$

$$= \Sigma_k\tau_k^2\left(\frac{P(E|DC_k)}{nP(\bar{E}C_k|D)}+\frac{P(E|\bar{D}C_k)}{mP(\bar{E}C_k|\bar{D})}\right) \tag{29}$$

Substitution of sample moments in (29) gives the consistent estimate v^2 of (13).

The estimate $\tilde{\rho} = 1 - A\tilde{R}$ obtained from the *matched design* depends on the data only through the vectors $\mathbf{x}_2, \mathbf{y}_2$ and the known control : case ratios $R_k, k = 1, \ldots, K$. Thus the linear expansion of $\tilde{\rho}$ appropriate for this design is given by (23) with $\mathbf{x}$ and $\boldsymbol{\pi}$ replaced by $\mathbf{x}_2$ and $\boldsymbol{\pi}_2$, and with $\boldsymbol{\rho}_{x_2} = \langle \tau_k/\pi_{2k} \rangle$ and $\boldsymbol{\rho}_{y_2} = \langle -\tau_k p_{\cdot k}/(R_k p_{2k}\pi_{\cdot k}) \rangle$. This gives

$$A \operatorname{var} \tilde{\rho} = n^{-2}(\boldsymbol{\rho}_{x_2}', \boldsymbol{\rho}_{y_2}')S_*(\boldsymbol{\rho}_{x_2}', \boldsymbol{\rho}_{y_2}')'$$

where S_* is the $2K \times 2K$ submatrix obtained by deleting the first K rows and columns of S in (25). Substituting sample moments in the resulting expression

$$A \operatorname{var} \tilde{\rho} = n^{-1}\Sigma_k\tau_k^2\left(\frac{\pi_{1k}}{\pi_{2k}}+\frac{p_{1k}}{R_kp_{2k}}\right)$$

$$= n^{-1}\Sigma_k\tau_k^2\left(\frac{P(E|DC_k)}{P(\bar{E}|DC_k)}+\frac{P(E|\bar{D}C_k)}{R_kP(\bar{E}|\bar{D}C_k)}\right) \tag{30}$$

again gives the consistent estimate v^2.

APPENDIX II

We prove that length of the ML interval is less than the length of the log interval whenever these intervals have non-zero length. It is evident from the interval definitions (16) and (18) that the desired result is equivalent to the inequality

$$\hat{\rho}[e^{w/\hat{\rho}} - e^{-w/\hat{\rho}}] - 2w > 0 \tag{31}$$

When $\hat{\rho} = 0$, expression (31) is infinite for all finite w. When $\hat{\rho} > 0$, the left-hand side of (31) can be rewritten

$$f(x) = e^x - e^{-x} - 2x$$

where $x = w/\hat{\rho} \geq 0$. It remains to show that $f(x) \geq 0$, with equality holding only in the trivial case $x = w = 0$. But this follows immediately from the fact that $f(0) = 0$ and that

$$f'(x) = e^x + e^{-x} - 2 = 2[\cosh x - 1]$$

which is positive for all $x > 0$.

APPENDIX III

For the random sampling design we obtained the bootstrap estimate (BSE) as follows:

1. We sampled n of the case observations with replacement. Thus we may have chosen one 'case' 3 times, while another not at all.
2. We sampled m of the control observations with replacement.
3. We used the bootstrap sample of cases and controls obtained in steps 1 and 2 in (9) to calculate attributable risk AR^*.
4. Steps 1–3 were independently repeated 200 times.
5. $BSE = \left[\frac{1}{199}\sum_{i=1}^{200}(AR^*_{(i)} - AR^*_{(\cdot)})^2\right]^{1/2}$ was calculated, where $AR^*_{(\cdot)}$ is the average of the 200 attributable risk values $AR^*_{(i)}$. For the matched design, we eliminated step 2 and modified step 1 so that each time we sampled a case observation, we selected as well the R control observations matched to the case.

REFERENCES

1. Levin, M. L. 'The occurrence of lung cancer in man', *Acta Un. Intern. Cancer*, **9**, 531–541 (1953).
2. Efron, B. 'Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods', *Biometrika*, **68**, 589–599 (1981).
3. Appel, M. F. 'Hereditary pancreatitis', *Arch. Surg.*, **108**, 63–65 (1974).
4. Hammond, E. C. 'Smoking in relation to the death rates of one million men and women', *Natl. Cancer Inst. Monogr. No. 19*, 127–204 (1966).
5. Walter, S. D. 'The estimation and interpretation of attributable risk in health research', *Biometrics*, **32**, 829–849 (1976).
6. Walter, S. D. 'The distribution of Levin's measure of attributable risk', *Biometrika*, **62**, 371–374 (1975).
7. Walter, S. D. 'Prevention for multifactorial diseases', *Am. J. Epidemiol.*, **112**, 409–416 (1980).
8. Leung, H. M. and Kupper, L. L. 'Comparisons of confidence intervals for attributable risk', *Biometrics*, **37**, 293–302 (1981).
9. Lundin, F. E., Wagoner, J. K. and Archer, V. E. 'Radon daughter exposure and respiratory cancer: quantitative and temporal aspects', *NIEHS and NIOSH Joint Monogr. No. 1* (1971).
10. Whittemore, A. S. and McMillan, A. 'Analyzing occupational cohort data: application to US uranium miners', in Prentice, R. L. and Whittemore, A. S. (Eds.) *Epidemiologic Methods – Risk Assessment* (in press).
11. *Vital Statistics of the United States, 1970 Volume II Mortality, Part A.*, National Center for Health Statistics, DHEW (HRA) 75–1101. Rockville, MD, 1974.
12. *Vital Statistics of the United States, 1977 Volume II Mortality, Part A.* USDHHS (PHS) No. 81–1101. US Govt Printing Office, Washington, DC, 1977.
13. Breslow, N. 'Odds ratio estimators when the data are sparse', *Biometrika*, **68**, 73–84 (1981).
14. Breslow, N. 'Design and analysis of case-control studies', *Annual Rev. Public Health*, **3**, 29–54 (1982).